

Semantic-Aware Depth Super-Resolution in Outdoor Scenes

Miaomiao Liu¹, Mathieu Salzmann², Xuming He¹

¹NICTA, ²EPFL

Abstract. While depth sensors are becoming increasingly popular, their spatial resolution often remains limited. Depth super-resolution therefore emerged as a solution to this problem. Despite much progress, state-of-the-art techniques suffer from two drawbacks: (i) they rely on the assumption that intensity edges coincide with depth discontinuities, which, unfortunately, is only true in controlled environments; and (ii) they typically exploit the availability of high-resolution training depth maps, which can often not be acquired in practice due to the sensors' limitations. By contrast, here, we introduce an approach to performing depth super-resolution in more challenging conditions, such as in outdoor scenes. To this end, we first propose to exploit semantic information to better constrain the super-resolution process. In particular, we design a co-sparse analysis model that learns filters from joint intensity, depth and semantic information. Furthermore, we show how low-resolution training depth maps can be employed in our learning strategy. We demonstrate the benefits of our approach over state-of-the-art depth super-resolution methods on two outdoor scene datasets.

1 Introduction

Depth sensors are becoming increasingly popular in many applications, such as virtual reality and autonomous navigation. While huge progress has been made in the development of such sensors, a typical example of which is the Kinect, for outdoor scenes, existing sensors remain limited in the spatial resolution they provide. As a consequence, depth super-resolution has emerged as a way to compute high-resolution depth maps from low-resolution ones.

In the past decade, many depth super-resolution methods have been proposed, such as filtering-based techniques [1,2], learning-based approaches [3,4] and CRF-based methods [5,6]. In particular, a popular trend consists of exploiting the high-resolution intensity image corresponding to the low-resolution depth map to constrain the depth super-resolution process [2,4]. The intuition behind such an approach is that the discontinuities in intensity space correspond to those in depth. Therefore, the high-resolution depth maps obtained in this manner should be less prone to over-smoothing. While the assumption of corresponding discontinuities is valid in nicely engineered environments, such as in the Middlebury dataset, which is the most popular benchmark for depth super-resolution algorithms, it typically does not hold in more realistic scenarios, and in particular, as illustrated in Fig. 1, in outdoor scenes, where disturbances such as strong shadows are common. More importantly, state-of-the-art algorithms typically assume to have access to high-resolution depth maps in a training stage to learn the

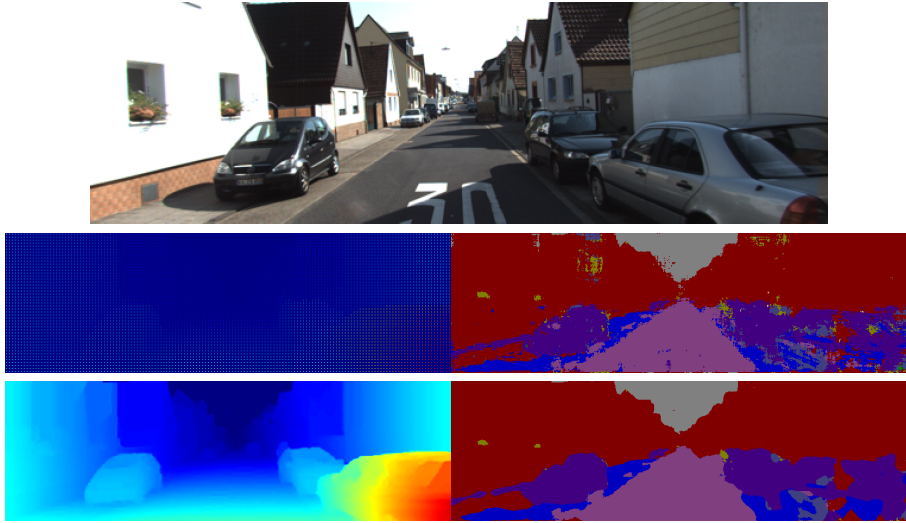


Fig. 1. Semantic-aware depth super-resolution in outdoor scenes. **Top:** color image. **Middle:** sparse disparity observations (left) and noisy semantics (right). **Bottom:** estimated high-resolution disparity (left) and improved semantics using our method (right). Best viewed in color.

operators they will apply at test time. Unfortunately, for outdoor scenes where depth sensors can only produce low-resolution depth maps, this kind of training data is unavailable.

In this paper, we introduce an approach to performing depth super-resolution in these challenging conditions. In particular, in addition to image gradient, we propose to rely on semantic information, in the form of pixel-wise image labelings, to better constrain the depth super-resolution process. Semantic maps do not directly suffer from lighting conditions, and transitions between objects often correspond to depth discontinuities. Furthermore, thanks to the popularity of semantic labeling in computer vision, many datasets with ground-truth semantic maps are available for outdoor scenes, and existing methods produce increasingly accurate predictions. In a training stage, our approach therefore jointly exploits intensity and ground-truth semantics, and, importantly, tackles the realistic scenario where only *low-resolution* training depth maps are available. At test time, given an image and low-resolution depth measurements, we generate a high-resolution depth map while simultaneously denoising a predicted semantic labeling.

More specifically, our approach relies on the co-sparse analysis model introduced in [7]. The analysis model learns operators, i.e., filters, such that the response of those filters is sparse when applied to valid data patches. In our case, during training, we therefore seek to reconstruct a high-resolution depth map, while simultaneously learning operators that generate sparse responses from intensity, depth and semantics patches. At test time, since we cannot assume to have access to ground-truth semantic labels, we make use of existing frameworks [8,9] to predict a semantic map. We then aim to

jointly denoise this semantic map and predict a high-resolution depth map from low-resolution measurements, such that the learned operators yield sparse responses when applied to the predicted semantics and depth, together with the corresponding intensity. This process can be written as a series of convolutions, which we implemented on a GPU.

We evaluated our approach on two challenging outdoor datasets: the KITTI benchmark [10] and Make3D [11]. Our experiments evidence the benefits of exploiting semantic information for depth super-resolution. Furthermore, they also show that our approach to learning operators from low-resolution depth maps comes at very little loss compared to training with high-resolution ones, which are typically unavailable in practice. Finally, as an additional benefit, our approach also yields improved semantic labelings.

2 Related Work

Densifying depth maps has attracted much attention in recent years due to the popularity of 3D sensors providing semi-dense measurements for both indoor and outdoor scenes. The resulting depth super-resolution approaches can be roughly categorized into two classes: the methods that rely on depth only, and those that also exploit other modalities. While a large body of work addresses the problem of multiview depth fusion (e.g., [12,13,14]), here, we focus on depth super-resolution from a single view.

The first kind of methods treats depth as a 2D image and adopts techniques from the image super-resolution literature. For instance, Mac Aodha et al. [15] infer high-resolution depth from a single low-resolution depth image based on a generic database of depth patches at high-resolution. Hornacek et al. [16] use the concept of self-similarity in 3D and super-resolve depth from a single low-resolution depth image. While effective in their context, these methods cannot leverage additional cues about the underlying scene, such as image intensity. In real-world applications, however, measurements from other modalities are typically available.

The second category of methods therefore emerged as a solution to exploit the correlation between two modalities, typically using high-resolution intensity images to guide depth super-resolution. For instance, Diebel and Thrun [5] use an image contrast aware pairwise MRF to super-resolve a low-resolution range image. Park et al. [6] improve this MRF model by incorporating outlier detection, richer image cues and smoothness priors over larger neighborhoods. In [17], color and depth images are jointly upsampled from low-resolution stereo images. In essence, however, these methods only capture pairwise intensity-sensitive depth smoothness, and thus lack the capacity to model higher-order depth patterns.

To explore longer-range dependencies, Yang et al. [2] perform depth super-resolution via a cross bilateral filtering operation on depth given the corresponding intensity image. Yu et al. [18] use a shape-from-shading model to refine the depth estimation. Shen and Cheung [19] propose a layered model for depth completion. More recently, Lu et al. [3] introduced a low-dimensional subspace model for RGB-D patches, which was used to jointly complete and denoise depth maps. Perhaps most related to our work is the co-sparse analysis model of [7], which learns a set of joint analysis operators in

intensity and depth to regularize depth super-resolution. By considering depth patches, this method can capture more complex patterns and enforce longer-range constraints on the depth map. The analysis model was also employed in [4], but by using pre-defined operators applied on depth only.

Despite this progress, however, all the existing works solely focus on depth completion in indoor scenes, such as Middlebury images [20], where the boundaries in intensity images are well aligned with depth discontinuities. Here, we consider the more challenging case of outdoor scenes, and introduce the use of semantics to handle large illumination changes and shadows in outdoor images. Furthermore, most learning-based approaches rely on high-resolution depth maps in the training stage, e.g., [15,7], which are typically challenging to obtain in outdoor scenes. By contrast, we design a learning algorithm that only requires low-resolution depth data.

While our main focus is depth super-resolution, our model also lets us improve noisy semantic labelings. The relationship between depth estimation and semantic labeling has been studied in several recent works. In particular, RGB-D images have been employed as input to semantic labeling algorithms [21,22,23]. More related to our goal, Liu et al. [24] use predicted semantic labels to estimate depth in outdoor scenes. Similarly, Ladický et al. [25] learn a joint classifier to predict semantic labels and depth values of image patches. Other approaches integrate depth reconstruction and semantic labeling using stereo images or image sequences as input [26,27]. Unlike our approach, however, these methods do not tackle the problem of depth super-resolution, and typically use high-resolution depth maps or 3D models in the training process.

3 Semantic-Aware Depth Super-Resolution

Our goal is to estimate a high-resolution depth map $\mathbf{D} \in \mathbb{R}^n$ from sparse and noisy measurements $\hat{\mathbf{D}} \in \mathbb{R}^m$, where, typically, $m \ll n$. This depth super-resolution process can be expressed as the problem of finding \mathbf{D} , such that

$$\hat{\mathbf{D}} = \mathcal{A}\mathbf{D} + \mathbf{e}_D, \quad (1)$$

where \mathcal{A} models the down-sampling process and $\mathbf{e}_D \in \mathbb{R}^m$ denotes the noise. Computing \mathbf{D} from Eq. 1 is an ill-posed problem since m is significantly smaller than n . Here, we aim to address this problem by incorporating both intensity and pixel-level semantics to regularize depth estimation.

In particular, we adopt the analysis model framework of [28] and introduce an approach to building a joint prior on image, depth and semantic patches. The analysis model captures the signal structure by learning an analysis operator $\mathbf{\Omega} \in \mathbb{R}^{k \times n}$, such that applying $\mathbf{\Omega}$ to the input signal yields a sparse output vector. Here, we design a trimodal co-sparse analysis approach to depth super-resolution in outdoor scenes by exploiting the strong correlations in the discontinuity patterns of multimodal cues, i.e., intensity, depth and semantics. Furthermore, we introduce a new learning method that prevents the requirement for high-resolution training depth maps, and thus better suits the outdoor setting in which most sensors simply cannot produce high-resolution depth. We first discuss our depth super-resolution framework in this section, and then present our learning approach in Section 4.

3.1 A Trimodal Co-Sparse Analysis Model

We now introduce our trimodal co-sparse analysis model. To this end, let us first consider the case of a single patch from a single modality, e.g., an image patch. Given an image patch \mathbf{X} (in vector form) and an operator Ω , the analysis model encodes the structure of the signal by the zero entries in the vector $\Omega\mathbf{X}$. In other words, the signal \mathbf{X} lies in the null space of the matrix composed by a subset of the rows of Ω . Specifically, this subset, defined as the *co-support* of \mathbf{X} , can be written as

$$\Gamma := \{j | (\Omega\mathbf{X})_j = 0\}, \quad (2)$$

where $(\cdot)_j$ denotes the j -th element of a vector. In our case, we aim to capture the patterns common to intensity, depth and semantic patches. To this end, let $\mathbf{I} \in \mathbb{R}^{n_I}$ and $\mathbf{D} \in \mathbb{R}^{n_D}$ denote a vectorized intensity patch and depth patch, respectively. For the semantics, we encode the class label of a pixel as an L -dimensional vector representing the probability for the pixel to belong to each of the L classes of interest. The semantic information of the pixels in a patch can therefore be grouped in a vector $\mathbf{S} \in \mathbb{R}^{Ln_S}$.

Let us then denote by $\Omega_I \in \mathbb{R}^{k \times n_I}$, $\Omega_D \in \mathbb{R}^{k \times n_D}$, and $\Omega_S \in \mathbb{R}^{k \times Ln_S}$ the operators corresponding to \mathbf{I} , \mathbf{D} and \mathbf{S} , respectively. Similarly, let Γ_I , Γ_D , and Γ_S be the co-support of \mathbf{I} , \mathbf{D} , and \mathbf{S} . Our model assumes that the structures of the different modalities are correlated, and thus that Γ_I , Γ_D , and Γ_S are statistically highly dependent. In other words, the probability for a row index j , $1 \leq j \leq k$, to be in the co-support of one modality should be higher if j also is in the co-support of the other modalities.

More concretely, for each row index j , the dependency between the co-supports of different modalities can be modeled by the function

$$g_j(\Omega_I\mathbf{I}, \Omega_D\mathbf{D}, \Omega_S\mathbf{S}) = \log \left(1 + \nu_I(\Omega_I\mathbf{I})_j^2 + \nu_D(\Omega_D\mathbf{D})_j^2 + \nu_S(\Omega_S\mathbf{S})_j^2 \right), \quad (3)$$

where ν_I , ν_D , and ν_S are the weights of the different modalities. This function, which was shown to be more effective than other sparsity-inducing functions such as the l_0 and l_1 norms [29], will be zero only if the three operators yield a zero value for index j , i.e., if j is in the co-support of all modalities. Therefore, the global cost function

$$g(\Omega_I\mathbf{I}, \Omega_D\mathbf{D}, \Omega_S\mathbf{S}) = \sum_{j=1}^k g_j(\Omega_I\mathbf{I}, \Omega_D\mathbf{D}, \Omega_S\mathbf{S}) \quad (4)$$

will reach its minimum if the co-supports of the three modalities coincide exactly.

While the prior induced by this cost favors discontinuities aligned in the three modalities, thanks to the weighted contributions of semantics and intensity, it still allows depth boundaries to occur between regions of the same semantic class. As such, and as discussed in the remainder of this paper, this cost function is therefore very-well suited to both perform depth super-resolution and learn the analysis operators. Below, we start by formulating the depth super-resolution problem.

3.2 Depth Super-Resolution with Co-Sparsity

We now turn to the problem of performing depth super-resolution for an entire image, given the co-sparse analysis operators Ω_I , Ω_D , Ω_S for the three modalities, i.e., in-

tensity, depth and semantics. Since these operators are defined on patches, we use the convolution operation to apply them to the entire image. We then define an energy that combines a regularizer based on our trimodal co-sparse analysis model with data terms accounting for the sparse depth measurements and the noisy semantic label predictions.

More specifically, let $\hat{\mathbf{D}}$ be the vector of sparse depth measurements given as input with corresponding intensity image \mathbf{I} . From \mathbf{I} , and given a pre-trained semantic classifier, we can predict a noisy semantic map $\hat{\mathbf{S}}$, as is commonly done in image-based semantic labeling. Given these two inputs, we seek to find the high-resolution depth map \mathbf{D} and noise-free semantic labels \mathbf{S} that minimize a data term of the form

$$E_d(\mathbf{D}, \mathbf{S}) = \|\mathbf{A}\mathbf{D} - \hat{\mathbf{D}}\|_2^2 + \lambda \|\mathbf{S} - \hat{\mathbf{S}}\|_2^2, \quad (5)$$

where λ encodes the relative importance of the two terms. This data term encourages the estimated depth to be close to the sparse observations, and the estimated semantic map to be close to the noisy one. On its own, of course, this data term is not sufficient since, for instance, nothing constrains the depth values with no measurements and nothing relates the depth to the semantics. We propose to make use of our trimodal co-sparse analysis model to address this issue.

As mentioned above, we convolve our patch-based analysis operators with the entire image. To this end, let us define position operators $\{\mathbf{P}_{ij}^I, \mathbf{P}_{ij}^D, \mathbf{P}_{ij}^S\}$ that extract vectorized local patches centered at pixel location (i, j) and of size $\sqrt{n_I} \times \sqrt{n_I}$, $\sqrt{n_D} \times \sqrt{n_D}$, $\sqrt{n_S} \times \sqrt{n_S}$ from \mathbf{I} , \mathbf{D} , and \mathbf{S} , respectively. Following the trimodal co-sparse analysis model of Section 3.1, we define a regularizer of the form

$$E_s(\mathbf{D}, \mathbf{S}|\mathbf{I}) = \frac{1}{n} \sum_{r=1}^n g(\Omega_I \mathbf{P}_r^I \mathbf{I}, \Omega_D \mathbf{P}_r^D \mathbf{D}, \Omega_S \mathbf{P}_r^S \mathbf{S}), \quad (6)$$

where n is the total number of pixels in the image. This regularizer now relates the estimated depth to the estimated semantics, and links them to the given intensity image.

By combining this regularizer with our data term, we formulate joint depth super-resolution and semantic labeling as the optimization problem

$$(\mathbf{D}^*, \mathbf{S}^*) = \underset{\{\mathbf{D} \in \mathbb{R}^n, \mathbf{S} \in \mathbb{R}^{Ln}\}}{\operatorname{argmin}} \quad \eta E_s(\mathbf{D}, \mathbf{S}|\mathbf{I}) + E_d(\mathbf{D}, \mathbf{S}), \quad (7)$$

where η is a weight that adjusts the influence of the regularizer. In practice, we make use of a conjugate gradient descent algorithm to solve (7).

4 Learning Multimodal Analysis Operators

The semantic-aware depth super-resolution method introduced in Section 3 relies on given analysis operators. We now turn to the problem of learning these operators from training data. In particular, we consider two different settings: one where we exploit high-resolution training depth maps, and a more realistic one where we only rely on low-resolution training depth. In both cases, we assume that the ground-truth depth, the intensity image and the semantic probability map are registered.

4.1 Learning with High-Resolution Depth

Given a set of high-resolution multimodal patches $\{\mathbf{I}_m, \mathbf{D}_m, \mathbf{S}_m\}_{m=1}^M$, we make use of the trimodal co-sparse analysis model to learn the analysis operators corresponding to the three modalities. Intuitively, we search for operators such that the co-supports of those three modalities are highly dependent. By making use of the cost defined in Eq. 4 for a single patch, we can express this as the loss

$$L_s(\mathbf{\Omega}_I, \mathbf{\Omega}_D, \mathbf{\Omega}_S) = \frac{1}{M} \sum_{m=1}^M g(\mathbf{\Omega}_I \mathbf{I}_m, \mathbf{\Omega}_D \mathbf{D}_m, \mathbf{\Omega}_S \mathbf{S}_m). \quad (8)$$

Minimizing this loss alone, however, suffers from trivial solutions (i.e., the operators can simply go to 0). To avoid these trivial solutions, we make use of the constraints and priors introduced in [28].

More specifically, as a first constraint, we assume that the (transposed) analysis operators lie on the oblique manifold. In other words, each operator $\mathbf{\Omega}_X$, where X can be either D , I , or S , is constrained as

$$\begin{aligned} \mathbf{\Omega}_X^T &\in OB(n, k) = \\ &\{\mathbf{\Omega}^T \in \mathbb{R}^{n \times k} | \text{rank}(\mathbf{\Omega}) = n, \text{ddiag}(\mathbf{\Omega}\mathbf{\Omega}^T) = \mathbf{I}_k\}, \end{aligned} \quad (9)$$

which indicates that $\mathbf{\Omega}_X$ must have full rank and unit-norm rows. Furthermore, we also make use of two additional priors introduced in [28] that have proven effective to learn meaningful operators. These priors can be expressed as

$$\begin{aligned} h(\mathbf{\Omega}_X) &= -\frac{1}{n \log(n)} \log \det \left(\frac{1}{k} \mathbf{\Omega}_X^T \mathbf{\Omega}_X \right), \\ r(\mathbf{\Omega}_X) &= -\sum_{1 \leq i < j \leq k} \log \left(1 - (\langle \omega_X^i, \omega_X^j \rangle)^2 \right), \end{aligned} \quad (10)$$

where ω_X^i is the i^{th} row of $\mathbf{\Omega}_X$, and $\langle \cdot \rangle$ denotes the inner product between two vectors. The function $h(\mathbf{\Omega})$ encodes a rank constraint, and $r(\mathbf{\Omega})$ encourages the rows of $\mathbf{\Omega}$ to be linearly independent. These priors can then be grouped in a regularizer of the form

$$\begin{aligned} L_c(\mathbf{\Omega}_I, \mathbf{\Omega}_D, \mathbf{\Omega}_S) &= \kappa_I h(\mathbf{\Omega}_I) + \kappa_D h(\mathbf{\Omega}_D) + \kappa_S h(\mathbf{\Omega}_S) \\ &\quad + \mu_I r(\mathbf{\Omega}_I) + \mu_D r(\mathbf{\Omega}_D) + \mu_S r(\mathbf{\Omega}_S), \end{aligned} \quad (11)$$

where $\kappa_{\{I,D,S\}}$ and $\mu_{\{I,D,S\}}$ are weights that adjust the influence of the individual terms.

By combining the loss of Eq. 8 and the regularizers discussed above, learning can be expressed as

$$\begin{aligned} \min_{\mathbf{\Omega}_{\{I,D,S\}}} & \eta L_s(\mathbf{\Omega}_I, \mathbf{\Omega}_D, \mathbf{\Omega}_S) + L_c(\mathbf{\Omega}_I, \mathbf{\Omega}_D, \mathbf{\Omega}_S) \\ \text{s.t. } & \mathbf{\Omega}_I^T \in OB(n_I, k), \mathbf{\Omega}_D^T \in OB(n_D, k), \mathbf{\Omega}_S^T \in OB(Ln_S, k), \end{aligned} \quad (12)$$

where η denotes the relative weight of the two terms. In practice, we use the geometric conjugate gradient descent method of [30] to solve this problem.

4.2 Learning with Low-Resolution Depth

In a more realistic scenario, such as for outdoor scenes, high-resolution depth maps are typically not available, even at training time. To tackle this scenario, we introduce an approach that exploits low-resolution depth information during training. More precisely, our training set is now comprised of registered high-resolution intensity patches, low-resolution depth patches and ground-truth semantic patches, which we denote by $\{\mathbf{I}_m, \hat{\mathbf{D}}_m, \mathbf{S}_m\}_{m=1}^M$.

We follow a similar intuition as in the high-resolution case, and search for operators that yield co-support sets which are highly dependent among the three modalities. To this end, we can re-use the loss of Eq. 8, as well as the regularizer of Eq. 11. Note, however, that the loss is defined on a complete depth map, which we do not have access to here. To address this issue, we propose to simultaneously estimate the high-resolution depth map \mathbf{D}_m of each training patch and the analysis operators.

To this end, we make use of a similar data term as for depth super-resolution. This data term encourages the reconstructed depth map to be consistent with the low-resolution depth map during the training process, and can thus be written as

$$L_d(\{\mathbf{D}_m\}) = \frac{1}{M} \sum_{m=1}^M \|\hat{\mathbf{D}}_m - \mathcal{A}\mathbf{D}_m\|_2^2, \quad (13)$$

where \mathcal{A} is the downsampling operator.

This lets us express learning with low-resolution depth maps as the optimization problem

$$\begin{aligned} \min_{\Omega_{\{I,D,S\}}, \{\mathbf{D}_m\}} & L(\Omega_I, \Omega_D, \Omega_S, \{\mathbf{D}_m\}) + L_d(\{\mathbf{D}_m\}) \\ \text{s.t. } & \Omega_I^T \in OB(n_I, k), \Omega_D^T \in OB(n_D, k), \Omega_S^T \in OB(Ln_S, k), \end{aligned} \quad (14)$$

where

$$L(\Omega_I, \Omega_D, \Omega_S, \{\mathbf{D}_m\}) = \eta L_s(\Omega_I, \Omega_D, \Omega_S, \{\mathbf{D}_m\}) + L_c(\Omega_I, \Omega_D, \Omega_S) \quad (15)$$

is now a function of the estimated depth.

Problem (14) has two different types of variables, i.e., the operators Ω_I, Ω_D and Ω_S that lie on the oblique manifold, and the depth maps $\{\mathbf{D}_m\}_{m=1}^M$ that are in Euclidean space. We therefore follow an alternating approach to computing these variables. In particular, we initialize $\{\mathbf{D}_m\}_{m=1}^M$ by interpolation of the low-resolution depth maps. As a first step, we fix the depth variables and optimize the operators by using geometric conjugate gradient descent on the manifold [30]. Then, we fix the resulting operators and optimize the depth variables using a conjugate gradient descent method in Euclidean space. We perform this alternating optimization for a fixed number of iterations. As evidenced by our experiments, this alternating strategy proved sufficient to achieve good results.

5 Experiments

We evaluated our approach on two challenging outdoor datasets: the KITTI benchmark [10] and Make3D [11]. To demonstrate the effectiveness of our method, we provide quantitative and qualitative evaluation results on both depth and semantics.

As mentioned in Section 3, we train the operators on patches. To this end, we extracted corresponding square patches of size $\sqrt{n_I} = \sqrt{n_D} = \sqrt{n_S} = 5$ from the intensity images, the depth maps and the semantic maps. The intensity and depth patches were then reshaped as 25-dimensional vectors. For the semantics, since we represent the label of each pixel as an L -dimensional vector encoding the probability of each label, vectorizing a semantic patch yields a $25 \cdot L$ -dimensional vector. We then subtracted the mean of each training patch, but did not normalize the patches. As is common practice with the analysis model, we learned redundant (or over-complete) operators, i.e., operators with more rows than the dimensionality of the vectorized patches. In particular, to trade-off accuracy and efficiency, we chose a factor 1.2 for the redundancy of Ω_S , which corresponds to the modality with highest dimensionality. We validated all the parameters of our model on the validation set of KITTI in a grid-search fashion. During training, we estimated the operators and the high-resolution depth maps by solving (14).

While we use the ground-truth semantics to learn the operators during training, we can only realistically have access to noisy semantic maps at test time. To generate such noisy semantic labels, we used existing multi-class classifiers [8,9]. We then represented the semantic information of each pixel as the classification confidence of each class. During test, we solved (7) to estimate the high-resolution depth and clean the noisy semantic map.

5.1 KITTI Dataset

The original KITTI dataset only provides the raw sparse disparity. Therefore, to quantitatively evaluate our method, we tested it on a subset of the KITTI data provided by Ladický et al. [25] and consisting of 60 images aligned with ground-truth dense disparity maps and semantic labels. We made use of the training and test splits provided with this data, namely, the first 30 images for training and the remaining ones for testing. We reserved 10 images from the original training set for validation purpose. We artificially created the low-resolution depth by downsampling the high-resolution ones by a factor $d \in \{2, 4, 8\}$ in both dimensions of the image. Therefore, our input for training includes the high-resolution images, the low-resolution depth maps and the ground-truth semantic labels. We used $L = 9$ classes in the dataset, as suggested in [25]. With our 1.2 redundancy factor, this yields $\Omega_S \in \mathbb{R}^{270 \times 225}$ (i.e., $5 \times 5 \times 9 = 225$). Since our trimodal co-sparse analysis model assumes that the operators all have the same number of rows, this yields $\Omega_I \in \mathbb{R}^{270 \times 25}$ and similarly $\Omega_D \in \mathbb{R}^{270 \times 25}$.

In the reconstruction process, the relative weight between the sparsity term and the semantic data term is fixed. However, we iteratively changed η and restarted the conjugate gradient descent ten times in order to reach a better local minimum. In particular, following the strategy of [7], we started with $\eta = 30$ and kept shrinking it to a final value $\eta = 0.04$. Our validation procedure resulted in the parameter values $\nu_I = \nu_D = 3$, $\nu_S = 30$ and $\lambda = \eta$.

As a first quantitative evaluation, we compare the results obtained with our approach when learning the operators with high-resolution depth with the following baselines: nearest-neighbor interpolation, bicubic interpolation, and the state-of-the-art *JID* [7], which also exploits high-resolution training depth maps. In Table 1, we report the root-mean-square-error (RMSE) of all the methods. Our approach yields lower RMSE than

method	2x	4x	8x
bicubic	1.6744	2.2466	3.4520
nearest-neighbour	2.0985	3.0034	4.1902
JID [7]	1.0169	1.5257	2.1921
ours-trainHR	0.9673	1.4906	2.1904

Table 1. Comparison of the results of our approach trained using high-resolution depth maps with three depth super-resolution baselines, including the state-of-the-art JID algorithm. We report the RMSE for three downsampling factors, i.e., 2, 4 and 8 times, respectively. These results evidence that depth super-resolution benefits from using semantic information.

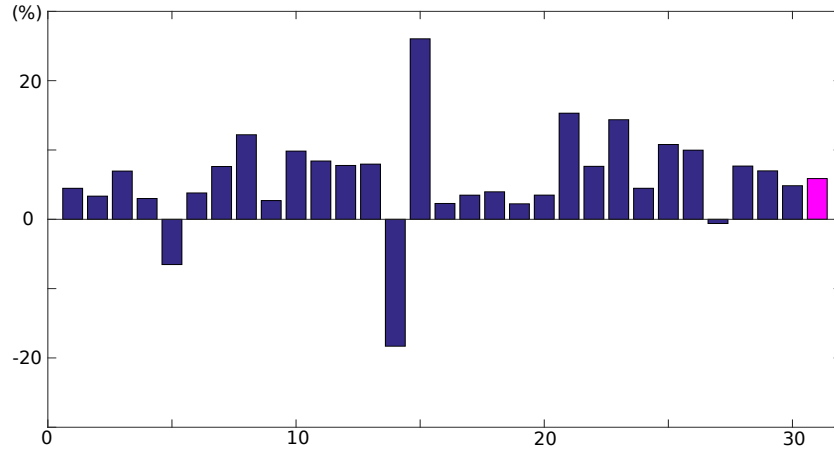


Fig. 2. Quantitative comparison of the results obtained by our method and of the JID results for each test image for 2x up-sampling. We show the relative improvement of our method over JID computed as $(JID_{RMSE} - Ours_{RMSE})/JID_{RMSE}$, where $Ours_{RMSE}$ and JID_{RMSE} denote the RMSE of our results and OF the JID results, respectively. Furthermore, we show the mean relative improvement in magenta, which is 5.8733%. These results evidence that, for most images, our improvement is significant.

all the baselines and of the JID results. In particular, it outperforms *JID*, which demonstrates the benefit of exploiting semantic information. In Fig. 2, we show the relative improvement of our method over JID, computed as $(JID_{RMSE} - Ours_{RMSE})/JID_{RMSE}$, where $Ours_{RMSE}$ and JID_{RMSE} denote the RMSE error of our results and of the JID results, respectively. Note that our approach yields a large improvement over JID on most images. Note also that the improvement achieved by our method is of similar magnitude to what is typically reported in the literature, e.g., [7].

In Fig. 3, we provide a qualitative comparison of our results with those of JID, which evidences that semantic information helps reducing the errors near object boundaries.

To show that our approach effectively exploits the semantic information, we re-ran the previous experiment with ground-truth semantics. This gave the following accuracies: 2x: 0.9418, 4x: 1.4743, 8x: 2.1590. This illustrates that better semantics can indeed

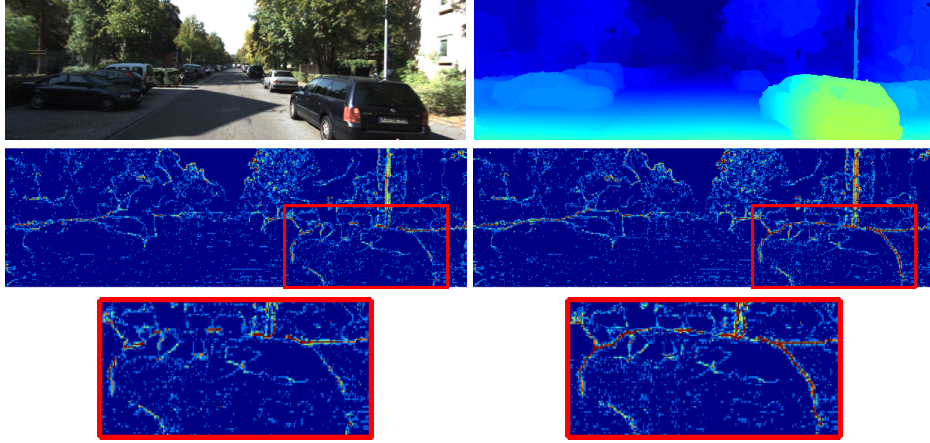


Fig. 3. Qualitative comparison of the results of our approach trained using high-resolution disparity with *JID*, which also relies on high-resolution training disparity maps. **Top:** RGB image and ground-truth disparity. **Middle:** Absolute difference between our results and ground-truth, and between the *JID* results and ground-truth, respectively. Blue denotes small errors and red large ones. **Bottom:** Close-up view of the portion highlighted in red. Note that our approach, which exploits semantics, yields lower errors near object boundaries (e.g., the car on the right). Best viewed in color.

further improve our results. Thus, as progress is made in semantic labeling, our method will produce increasingly accurate high-resolution depth maps.

method	2x	4x	8x
ours-trainLR	1.0546	1.5205	2.1861
ours-trainHR	0.9673	1.4906	2.1904

Table 2. Comparison of the results obtained by learning with high-resolution depth maps (ours-trainHR) and with low-resolution ones (ours-trainLR). Note that learning with low-resolution depth maps yields very little loss in accuracy.

In Table 2, we compare the results of our approach trained with low-resolution depth maps with those obtained by training with high-resolution ones. Note that training on low-resolution depth maps comes at very little loss in accuracy. Importantly, to the best of our knowledge, this realistic scenario has never been demonstrated in the past.

Since our method also denoises the semantic labels, we report the average per pixel and average per class accuracies of semantic labeling. As a first experiment, we used the multi-class classifier of [8] to generate noisy labels, which yields average per pixel and per class accuracies of 77.26% and 60.71%, respectively. After denoising with our approach, these accuracies increased to 81.19% and 63.98%. This is comparable to the state-of-the-art MRF model of [9], which yields accuracies of 81.82% and 64.31%. As a second experiment, we then started from this slightly more accurate result. After our

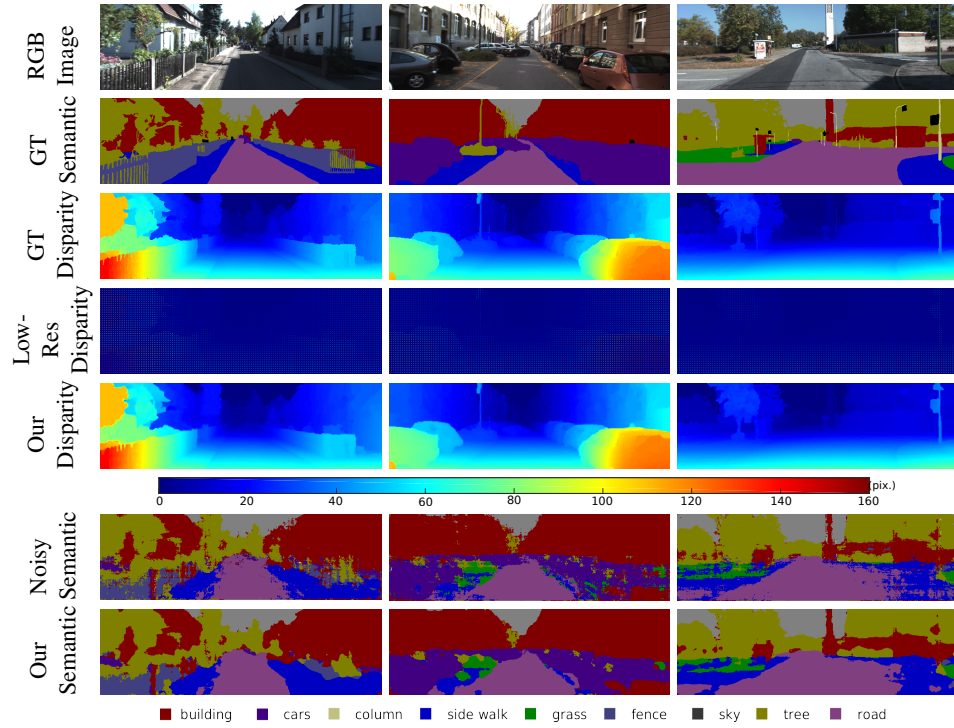


Fig. 4. Qualitative results on the KITTI dataset for a downsampling factor of 4. For the disparity values, red denotes large values, i.e., points close to the camera, and blue denotes small disparity values, i.e., points far from the camera. **From top to bottom:** RGB image, ground-truth semantics, ground-truth dense disparity map, the 6.25% observations mapped on the grid of the high-resolution image, our estimated high-resolution disparity map, noisy semantics predicted by the method of [8], our estimated semantics. Our method yields accurate high-resolution depth maps, and improves the semantic labeling results. Best viewed in color.

denoising process, the accuracies were further improved to 82.32% and 64.81%. We observed that these results are not sensitive to the sparsity of the depth map. In Fig. 4, we provide a qualitative evaluation of our results.

5.2 Make3d Dataset

We further evaluated our model on the Make3d dataset, another challenging outdoor dataset with sparse depth measurements. The ground-truth semantic labels for this dataset were provided by Liu et al. [24]. We made use of the training and test splits provided with their data, i.e., 400 training images and 134 test images. We estimated depth maps that match the size of the semantic images, i.e., 320×240 . Due to the lack of information about camera calibration, we obtained the observation mask by approximately mapping the low-resolution depth maps to the image grid, and excluding the

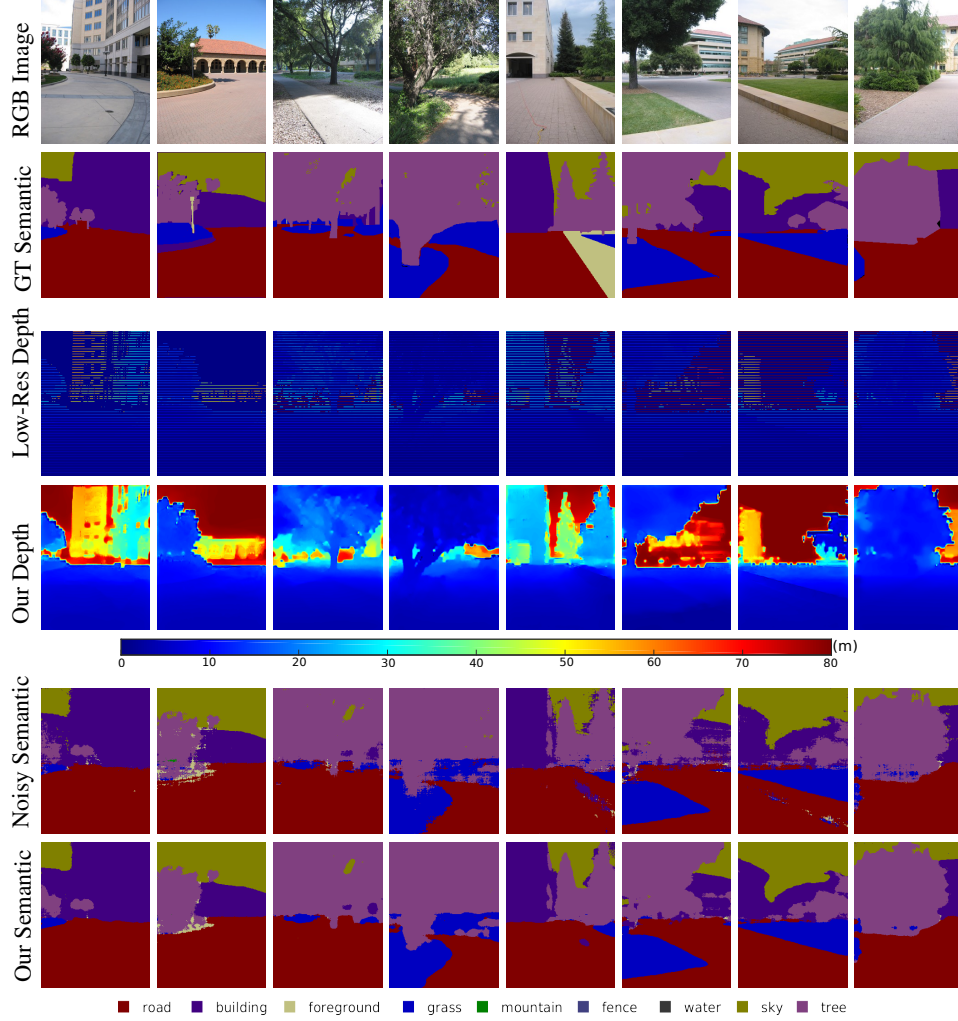


Fig. 5. Qualitative results on the Make3D dataset. From top to bottom: RGB image, ground-truth semantics, observed sparse depth map, estimated high-resolution depth map, noisy semantics obtained from [8] and our improved semantic labels. Best viewed in color.

pixels whose depth was greater than 78. This yields about 15% of observations on average. We then generated the noisy semantics using [8], which gave average per pixel and per class accuracies of 87.89% and 73.75%.

As before, we extracted square patches of size 5×5 on the training images, depth maps and semantic maps. This dataset contains $L = 8$ classes. With our 1.2 redundancy factor, this yields $\Omega_I \in \mathbb{R}^{240 \times 25}$, $\Omega_D \in \mathbb{R}^{240 \times 25}$, and $\Omega_S \in \mathbb{R}^{240 \times 200}$. We validated the parameters of our approach and obtained $\nu_S = 300$, $\nu_I = \nu_D = 3$. We used the

same optimization strategy as before, consisting of iteratively restarting the conjugate gradient descent, and started with $\eta = 300$ decreased to a final value of $\eta = 50$. Our method improves the semantic labeling accuracy to 88.67% and 73.81%. Since no ground-truth depth maps are available for this dataset, we can only perform a qualitative evaluation of our high-resolution depth maps. Some examples of these depth maps are provided in Fig. 5. Note that they look realistic and respect the object boundaries.

6 Conclusion

In this paper, we have presented a novel approach to depth super-resolution in the challenging outdoor setting, where intensity images have large variations due to illumination changes and shadows, and where high-resolution depth maps are difficult to acquire for training. In particular, we have proposed to incorporate semantic information into the super-resolution process, and have shown how to exploit low-resolution training depth maps. Our empirical evaluation on two outdoor datasets has demonstrated the effectiveness of our approach at predicting accurate high-resolution depth maps and semantic labelings. Furthermore, by outperforming state-of-the-art techniques, we have evidenced the benefits of exploiting semantics for depth super-resolution. In our current implementation, partially exploiting a GPU, reconstructing a high-resolution depth map takes 5 minutes for 500 iterations. In the future, we plan to speed this up by making better use of the GPU power. Furthermore, we intend to develop more effective methods to learn analysis operators.

References

1. Qi, F., Han, J., Wang, P., Shi, G., Li, F.: Structure guided fusion for depth map inpainting. *Pattern Recognition Letters* **34** (2013)
2. Yang, Q., Yang, R., Davis, J., Nistér, D.: Spatial-depth super resolution for range images. In: *CVPR*. (2007)
3. Lu, S., Ren, X., Liu, F.: Depth enhancement via low-rank matrix completion. In: *CVPR*. (2014)
4. Gong, X., Ren, J., Lai, B., Yan, C., Qian, H.: Guided depth upsampling via a cosparse analysis model. In: *CVPR Workshop on Multi-Sensor Fusion for Outdoor Dynamic Scene Understanding*. (2014)
5. Diebel, J., Thrun, S.: An application of markov random fields to range sensing. In: *NIPS*. (2005)
6. Park, J., Kim, H., Tai, Y.W., Brown, M., Kweon, I.: High quality depth map upsampling for 3d-tof cameras. In: *ICCV*. (2011)
7. Kiechle, M., Hawe, S., Kleinsteuber, M.: A joint intensity and depth co-sparse analysis model for depth map super-resolution. In: *ICCV*. (2013)
8. Gould, S.: DARWIN: A framework for machine learning and computer vision research and development. *Journal of Machine Learning Research (JMLR)* (2012)
9. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: *NIPS*. (2011)
10. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: *CVPR*. (2012)
11. Saxena, A., Chung, S.H., Ng, A.Y.: Learning depth from single monocular images. In: *NIPS*. (2005)

12. Schuon, S., Theobalt, C., Davis, J., Thrun, S.: Lidarboost: Depth superresolution for tof 3d shape scanning. In: CVPR. (2009)
13. Dolson, J., Baek, J., Plagemann, C., Thrun, S.: Upsampling range data in dynamic environments. In: CVPR. (2010)
14. Castaneda, V., Mateus, D., Navab, N.: Stereo time-of-flight. In: ICCV. (2011)
15. Mac Aodha, O., Campbell, N.D., Nair, A., Brostow, G.: Patch based synthesis for single depth image super-resolution. In: ECCV. (2012)
16. Hornáček, M., Rhemann, C., Gelautz, M., Rother, C.: Depth super resolution by rigid body self-similarity in 3d. In: CVPR. (2013)
17. Wang, L., Jin, H., Yang, R., Gong, M.: Stereoscopic inpainting: Joint color and depth completion from stereo images. In: CVPR. (2008)
18. Yu, L.F., Yeung, S.K., Tai, Y.W., Lin, S.: Shading-based shape refinement of rgb-d images. In: CVPR. (2013)
19. Shen, J., Cheung, S.: Layer depth denoising and completion for structured-light rgb-d cameras. In: CVPR. (2013)
20. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. IJCV (2002)
21. Koppula, H.S., Anand, A., Joachims, T., Saxena, A.: Semantic labeling of 3d point clouds for indoor scenes. In: NIPS. (2011)
22. Ren, X., Bo, L., Fox, D.: Rgb-(d) scene labeling: Features and algorithms. In: CVPR. (2012)
23. Gupta, S., Arbelaez, P., Malik, J.: Perceptual organization and recognition of indoor scenes from rgb-d images. In: CVPR. (2013)
24. Liu, B., Gould, S., Koller, D.: Single image septh estimation from predicted semantic labels. In: CVPR. (2010)
25. Ladický, L., Shi, J., Pollefeys, M.: Pulling things out of perspective. In: CVPR. (2014)
26. Ladický, L., Sturgess, P., Russell, C., Sengupta, S., Bastanlar, Y., Clocksin, W., Torr, P.H.: Joint optimization for object class segmentation and dense stereo reconstruction. IJCV (2012)
27. Haene, C., Zach, C., Cohen, A., Angst, R., Pollefeys, M.: Joint 3d scene reconstruction and class segmentation. In: CVPR. (2013)
28. Hawe, S., Kleinstüber, M., Diepold, K.: Analysis operator learning and its application to image reconstruction. IEEE TIP (2013)
29. Chen, Y., Ranftl, R., Pock, T.: Insights into analysis operator learning: From patch-based sparse models to higher order mrfs. IEEE TIP (2014)
30. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. (2008)